Predicting the Volume of Comments on Online News Stories

Manos Tsagkias e.tsagkias@uva.nl
Wouter Weerkamp w.weerkamp@uva.nl
Maarten de Rijke mdr@science.uva.nl

ISLA, University of Amsterdam
Science Park 107, 1098 XG Amsterdam

ABSTRACT

On-line news agents provide commenting facilities for readers to express their views with regard to news stories. The number of user supplied comments on a news article may be indicative of its importance or impact. We report on exploratory work that predicts the comment volume of news articles prior to publication using five feature sets. We address the prediction task as a two stage classification task: a binary classification identifies articles with the potential to receive comments, and a second binary classification receives the output from the first step to label articles "low" or "high" comment volume. The results show solid performance for the former task, while performance degrades for the latter.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Comment volume, prediction, feature engineering

1. INTRODUCTION

As we increasingly live our life online, in the form of blogs, discussion forums, comment facilities, etc., new types of data become available that can be mined for valuable knowledge. E.g., online chatter can be used to predict sales ranks of books [4]. Online news is an especially interesting data type for mining and analysis purposes. Much of what goes on in social media is a response to news events, as is evidenced by the large amount of news-related queries users submit to blog search engines [9]. Tracking news events and their impact as reflected in social media has become an important activity of media analysts [1]. We focus on online news articles plus the comments they generate, and attempt to predict news article comment volume prior to publication time.

One might raise the question why one should be interested in commenting behavior and the factors contributing to it. We envisage three types of application for predicting the volume of comments generated by news articles. First, *media and reputation analysis* is dependent on what users think of topics covered in the media. Predicting the comment volume might help in determining the desirability of an article (e.g., regarding the influence on one's reputation) or the timing of its publication (e.g., generate publicity and discussion during election time). Second, *pricing of news articles* by news agencies and *ad placement strategies* by news publishers could be made dependent on the expected comment volume; articles that are more likely to generate comments could be priced differently. Finally, news consumers could be served only news articles that are most likely to generate many comments; news sources can thus provide new services to their customers and can *save consumers' time* in identifying "important" articles.

Our aim in this paper is to predict comment volume of news articles prior to publication. To this end, we seek to answer the following two questions: (i) What are the dynamics of user generated comments on news articles? We look at article and comment statistics per source. (ii) Can we predict, prior to publication, whether a news story will receive any comments at all, and if so, whether it will receive few or many comments?

This work makes several contributions. First, it explores the dynamics of user generated comments in on-line Dutch media. Second, it introduces the problem of predicting the comment volume of a news article. Third, it provides a set of surface, cumulative, textual, semantic, and real-world features that can be used to predict the number of comments of a news story prior to publication. Fourth, it provides an evaluation of the introduced features. Fifth, an error analysis identifies possible causes for classification failure.

Section 2 contains related work; we explore news comments in Section 3; our feature sets are introduced in Section 4; predicting comment volume is done in Section 5; Section 6 contains discussion, error analyses, conclusions, and future work.

2. RELATED WORK

Different aspects of the comment space dynamics have been explored in the past. Schuth et al. [11] explore the news comments space of four on-line Dutch media, while Mishne and Glance [10] explored the weblog comment space. Kaltenbrunner et al. [6] measured community response time in terms of comment activity on Slashdot stories, and discovered regular temporal patterns on people's commenting behaviour. Recently, various prediction tasks and correlation studies have been considered in social media. Mishne and de Rijke [8] use textual features as well as temporal metadata of blog posts to predict the mood of the blogosphere. De Choudhury et al. [3] correlate blog dynamics with stock market activity, and Gruhl et al. [4] perform a similar task with blogs/reviews and book sales. Szabó and Huberman [12] predict the popularity of a

CIKM'09, November 2–6, 2009, Hong Kong, China.

story or a video on Digg or YouTube, given an item's statistics over a certain time period after publication. Lerman et al. [7] forecast the public opinion of political candidates from objective news articles. Finally, Tsagkias et al. [13] predict podcast preference using surface features extracted from podcast RSS feeds.

To our knowledge, no prediction tasks have been published that concern the volume of comments generated by online news articles.

3. EXPLORING NEWS COMMENTS

Our data consists of the aggregated content from seven on-line news agents: *Algemeen Dagblad* (*AD*), *De Pers*, *Financieel Dagblad* (*FD*), *Spits*, *Telegraaf*, *Trouw*, *WaarMaarRaar* (*WMR*), and one collaborative news platform, *NUjij*. We have chosen to include sources that provide commenting facilities for news stories, but differ in coverage (regional/national), in political views, in subject (general/politics/arts/entertainment), and in type. Six of the selected news agents publish daily newspapers and two, *WMR* and *NUjij*, are present only on the web. *WMR* publishes "oddly-enough" news and is interesting for observing the commenting behavior in this setting. *NUjij* is a collaborative news platform, similar to Digg, where people can submit links to news stories for others to vote for or start discussions on. We aggregate content for the period Nov 2008–Apr 2009, leaving us with a dataset of 290 375 articles, and 1 894 925 comments for all news sources.

News agent	Articles (commented)		Comments	Time (hrs) 0–1 com.	Time (hrs) 1–last com.
AD	41 740	(40%)	90 084	9.4	4.6
De Pers	61 079	(27%)	80 724	5.9	8.4
FD	9 911	(15%)	4 413	10.0	9.3
NUjij	94 983	(43%)	602 144	3.1	6.3
Spits	9 281	(96%)	427 268	1.1	13.7
Telegraaf	40 287	(21%)	584 191	2.5	30.2
Trouw	30 652	(8%)	19 339	11.7	8.1
WMR	2 442	(100%)	86 762	1.1	54.2

Table 1: Statistics of seven on-line news agents, and one collaborative news platform for the period Nov 2008–Apr 2009.

We turn to our first research question: What are the dynamics of user generated comments on news articles? Table 1 reports article and comment statistics per source, and the time between publication and the first comment, and between the first and last comment.

The volume of published articles per source varies per source. At one end we find large news sites such as *AD*, *De Pers*, *Telegraaf*, and *Trouw* with more than 30 000 published articles; the other end consists of smaller news agents, such as *FD*, *Spits*, and *WMR* with less than 10 000 published articles. We observe similar variation in the ratio of commented news articles. In general it is two times higher compared to the ratio of commented blog posts [10]. *Spits* and *WMR* find almost all of their articles commented on, while the ratio drops to lower than 10% for *Trouw*. We notice that *Spits* allows comments from guests, saving users from the registration process, and even though *WMR* allows comments only from registered users, the registration form needs minimal input, and is conveniently located just below the comment section. For *Trouw* comments are enabled only for some articles, partially explaining the low number of commented articles.

The elapsed time between an article's publication and its first comment is longer (6.7 hrs) compared to blogs (2.1 hrs) [10]. For some sources, comments start to arrive in the first two hours after article publication (e.g., *Spits* and *WMR*), while others receive tardy arrivals up to 10 hours after publication (e.g., *FD* and *Trouw*). Similar patterns govern the reaction lifetime, the time between the first and last comment.

Feature	Description	Type
Surface features		
month	Month (1-12)	Nom
wom	Week of the month (1-4)	Nom
dow	Day of the week (1-7)	Nom
day	Day of the month (1-31)	Nom
hour	Hour of the day (0-23)	Nom
first_half_hour	Publication in the first 30 minutes of the hour	Nom
art_char_length	Article content length	Int
category_count	Number of categories it is published on	Int
has_summary	Article has summary	Int
has_content	Article has content (HTML incl.)	Int
has_content_clean	Article has content (only text)	Int
links_cnt	Number of out-links	Int
authors_cnt	Number of authors	Int
Cumulative features		
art_same_hr	Published articles in same hour for source	Int
dupes_int_cnt	Near-duplicates in same source	Int
dupes_ext_cnt	Near-duplicates in other sources	Int
Textual features		
	tf of top-100 terms ranked by their log-likelihood score for each source	Int
Semantic features		
ne_loc_cnt	Number of location-type entities	Int
ne_per_cnt	Number of person-type entities	Int
ne_org_cnt	Number of organisation-type entities	Int
ne_misc_cnt	Number of miscellaneous-type entities	Int
has_local	Any entities referring to the Netherlands	Int
	tf of top-50 entities from each entity type, ranked by their log-likelihood score for each source	Int
Real-word features		
temperature	Temperature in Celsius at publication time	Num

Table 2: Listing of extracted features. The feature type is either nominal (nom), integer (int) or numeric (num).

The number of articles, the number of commented articles, the total number of comments, and the reaction times seem to be inherent characteristics of each source, possibly reflecting the credibility of the news organization, the interactive features they provide on their web sites, and their readers' demographics [2]. Our features attempt to capture the differences between the sources into account.

4. FEATURE ENGINEERING

We consider five groups of features: *surface*, *cumulative*, *textual*, *semantic*, and *real-world*. Table 2 summarizes all features.

Surface features. Feed metadata quality plays an important role in a user's decision to click a news item for reading or commenting. For example, if a news source supplies only the title of an article, but not a short summary, a user may prefer to click on a similar article from a different source that exposes more information.

Cumulative features. News agents broaden their news coverage through (inter)national news providers: A newsworthy story originating from a provider will therefore be published by multiple agents. The number of times we encounter a story in a time window is a good signal for it being interesting for multiple groups of readers and its exposure in multiple feeds increases its likelihood to be commented. On top of this, if the news supply is high, articles that could be commented, may not receive any comments because of the users' fast attention shift. We encode competition for attention by recording the number of published articles in the same hour.

Textual features. To collect textual features, or term sets, for each agent, we take the top-100 most discriminative unique terms using log-likelihood scores. Discriminative terms indicate differences between news sources. General news sources like *AD* and *Trouw*

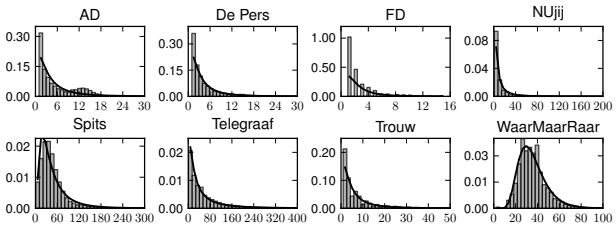

Figure 1: Modeling comment distribution per source using the continuous log-normal distribution (black line). The grey bars represent the observed data. y-axis stands for probability density, and x-axis stands for number of comments (binned).

show mainly "general" news terms (e.g., *Israeli* and *Palestinian*), while a financial news source (*FD*) has a clear preference for financial terms: *banks* and *Euro zone*. The online news sources (*NUjij* and *WMR*) differ from sources with an offline presence, preferring terms like *police*, *casino*, and *soccer*.

Semantic features. We apply named entity recognition to extract *persons*, *locations*, *organizations*, and *miscellaneous* entities. Similar to the discriminative terms, we select the top-50 most discriminative entities for each entity type. Politicians are popular entities for discussions: *Geert Wilders* (right-winged politician) and *Jan-Peter Balkenende* (Prime Minister) are among the ones attracting many comments. As to organizations, soccer clubs attract much discussion: *Ajax*, *PSV*, and *Feyenoord* are the three biggest soccer clubs in the Netherlands. Here again, politics is a popular topic: *Hamas*, *PvdA*, and *PVV* are dominant political organizations.

Real-world features. The last set of features explores the potential correlation between real-world environmental conditions (e.g., weather conditions) and commenting behavior. Here, for each article's publication time, we assign the median temperature in the Netherlands at that time as an indicator of good or bad weather.

5. PREDICTING COMMENT VOLUME

We now turn to the second research question: Can we predict, prior to publication, whether a news story will receive any comments at all? And if it receives comments, can we predict whether it receives few or many comments? Recall that for each news agent the comment volume may vary substantially. Before defining volume levels, the comment volume needs to be normalized across sources; we fit a log-normal distribution to each source, similar to the modeling of response time in Slashdot [5], and define the threshold between "low" and "high" volume at the inverse cumulative log-normal distribution function at 0.5 (see Figure 1).[1]

We address the prediction task as two consecutive classification tasks to compensate for the highly skewed datasets. First, we segregate articles with regard to their potential of receiving comments. A binary classification is performed with two classes: *with comments* vs. *without comments*. Second, we predict the comment volume level for the articles predicted to receive comments in the first step (positive class). This second classification is performed with two classes: *low volume* and *high volume*. We are not interested in optimizing classification performance, but rather in investigating whether different types of features can distinguish articles that hold potential to receive comments, and ultimately to quantify and predict this potential in terms of comment volume levels.

[1]Threshold for log-normal: *AD*: 3, *De Pers*: 3, *FD*: 2, *NUjij*: 6, *Spits*: 36, *Telegraaf*: 32, *Trouw*: 4, *WMR*: 34

5.1 Experimental set-up

We report on classification experiments per news source on the following experimental conditions: a baseline, one group of features at a time, and combining all feature groups. The baseline consists of six temporal features (*month*, *week of the month*, *day of the week*, *day of the month*, *hour*, and *first half hour*). For each source in our dataset, we create training and test sets. The training sets contain articles published from Nov 2008 until Feb 2009, and the test sets consist of the articles published in Mar 2009. We use RandomForest, a decision tree meta classifier. For evaluation of the classification performance we report the F1-score, and the percentage of correctly classified instances for each experimental condition. Significance of results is measured with the Kappa-statistic.

5.2 Stage 1: Any comments?

Looking at Table 3, most sources show a high F1 for the negative class, while only two sources show a high F1 for the positive class. These results reflect the commented/non-commented ratio of articles in each source that leads to highly skewed training sets. *WMR* and *Spits*, most of their articles having at least one comment, show a high ratio of positive examples, pushing the F1 score close to 1. As a result, for this classification experiment, the different groups of features are not expected to differ greatly for these two sources.

The baseline displays solid performance across the board. However, the Kappa-statistic hovers near zero, suggesting that if we classified the articles randomly, there is chance of observing similar results. Among the groups of features, textual and semantic features perform the best for most sources. This confirms that certain words and named entities trigger comments. Cumulative, surface, and real-world features perform similar to the baseline. Interestingly, the real-world features for *AD* achieve an F1 score of 0.749 for the negative class with Kappa at 0.48), and the surface features' performance for *Trouw* has an F1 score of 0.952 for the negative class with Kappa at 0.36. The combination of all groups of features does not lead to substantial improvements, but hovers at similar levels when using textual features only.

5.3 Stage 2: High vs. low volume

For the second classification experiment, articles that have previously been classified as *yes comments* are now classified based on whether they will receive a *high* or *low volume* of comments. Misclassified negative examples (articles without comments) from the first stage are labeled *low volume*. Five sources lack results for the real-world feature set due to the classifier marking all articles as negative in the first step.

In this setting, the F1 score is more equally distributed between the negative and the positive class. Textual and semantic features prove again to be good performers with non-zero Kappa, although varying almost 24% between sources (*NUjij* vs. *FD*). The variation suggests that the number of textual and semantic features should be optimized per source. The performance of cumulative features varies substantially between sources. E.g., for *Trouw* and *NUjij* it is among the best performing groups, but for *FD* it has a negative Kappa index. Looking at all groups combined, Kappa values increase, an indication for more robust classification. In general, the classification performance for all groups combined is better than the baseline, although the difference depends on the source. Comparing the performance of all features and individual feature sets, we observe that in some cases performance degrades in favor of a higher Kappa-value: For *Telegraaf* for example, textual features alone classify 55% of the instances correct (Kappa: 0.06), while all features reach 51% correctly classified instances (Kappa: 0.14).

Feature group	*yes/no comments* F1 (N)	F1 (Y)	K	Acc.	*low/high volume* F1 (L)	F1 (H)	K	Acc.
Soure: ***AD***								
Baseline	0.70	0.29	0.04	58%	0.39	0.40	0.00	40%
Surface	0.67	0.38	0.06	57%	0.49	0.39	0.02	45%
Cumulative	0.74	0.14	0.03	60%	0.44	0.49	0.07	48%
Textual	0.73	0.43	0.19	64%	0.45	0.54	0.09	50%
Semantic	0.72	0.37	0.14	62%	0.51	0.48	0.05	50%
Real-world	0.75	0.00	0.48	60%				
All	0.73	0.41	0.16	63%	0.54	0.51	0.11	53%
Soure: ***De Pers***								
Baseline	0.82	0.00	0.00	69%				
Surface	0.81	0.01	0.00	68%	0.69	0.36	0.12	58%
Cumulative	0.81	0.12	0.04	68%	0.48	0.34	-0.03	42%
Textual	0.81	0.35	0.19	70%	0.65	0.52	0.19	59%
Semantic	0.80	0.33	0.17	69%	0.62	0.48	0.15	56%
Real-world	0.82	0.00	0.00	69%				
All	0.82	0.27	0.15	71%	0.61	0.58	0.20	59%
Soure: ***FD***								
Baseline	0.91	0.07	0.03	84%	0.28	0.28	0.01	28%
Surface	0.91	0.22	0.16	84%	0.42	0.53	0.09	48%
Cumulative	0.91	0.05	0.02	84%	0.49	0.08	-0.19	34%
Textual	0.91	0.40	0.32	85%	0.42	0.53	0.09	48%
Semantic	0.92	0.21	0.16	85%	0.35	0.50	0.00	44%
Real-world	0.92	0.00	0.00	85%	0.55	0.52	0.14	53%
All	0.92	0.25	0.19	85%	0.52	0.66	0.25	60%
Soure: ***NUjij***								
Surface	0.60	0.21	0.02	47%	0.68	0.35	0.10	57%
Cumulative	0.56	0.30	0.00	46%	0.80	0.32	0.12	69%
Textual	0.63	0.59	0.24	61%	0.70	0.57	0.28	65%
Semantic	0.59	0.55	0.17	57%	0.75	0.53	0.29	68%
Real-world	0.61	0.00	0.0	44%				
All	0.65	0.40	0.17	56%	0.62	0.66	0.28	64%
Soure: ***Spits***								
Baseline	0.00	0.99	0.00	99%	0.38	0.67	0.10	57%
Surface	0.08	0.99	0.08	99%	0.42	0.69	0.11	59%
Cumulative	0.00	0.99	0.00	99%	0.27	0.74	0.04	61%
Textual	0.00	0.99	0.00	98%	0.50	0.56	0.11	53%
Semantic	0.00	0.99	0.00	98%	0.40	0.66	0.06	56%
Real-world	0.00	0.99	0.00	99%	0.13	0.77	0.00	63%
All	0.00	0.99	0.00	99%	0.48	0.64	0.13	57%
Soure: ***Telegraaf***								
Baseline	0.89	0.12	0.07	80%	0.43	0.28	0.00	37%
Surface	0.88	0.12	0.06	79%	0.50	0.31	0.00	42%
Cumulative	0.89	0.00	0.00	80%	0.25	0.40	0.07	33%
Textual	0.87	0.26	0.14	78%	0.66	0.36	0.06	55%
Semantic	0.87	0.19	0.10	78%	0.58	0.35	0.07	49%
Real-world	0.89	0.00	0.00	80%				
All	0.89	0.17	0.11	80%	0.51	0.51	0.14	51%
Soure: ***Trouw***								
Baseline	0.95	0.11	0.10	90%	0.38	0.22	-0.4	31%
Surface	0.95	0.29	0.36	91%	0.44	0.48	-0.06	46%
Cumulative	0.95	0.02	0.01	90%	0.55	0.44	0.14	50%
Textual	0.96	0.63	0.59	93%	0.42	0.54	0.01	49%
Semantic	0.95	0.37	0.33	91%	0.49	0.55	0.09	52%
Real-world	0.95	0.00	0.15	90%				
All	0.96	0.54	0.50	93%	0.44	0.56	0.04	51%
Soure: ***WMR***								
Baseline	0.00	1.00	1.00	100%	0.45	0.51	0.10	48%
Surface	0.00	1.00	1.00	100%	0.44	0.50	0.03	47%
Cumulative	0.00	1.00	1.00	100%	0.47	0.01	-0.01	31%
Textual	0.00	1.00	1.00	100%	0.48	0.54	0.10	51%
Semantic	0.00	1.00	1.00	100%	0.43	0.53	0.06	52%
Real-world	0.00	1.00	1.00	100%	0.48	0.00	0.00	31%
All	0.00	1.00	1.00	100%	0.45	0.54	0.06	50%

Table 3: Binary classification of articles into articles with (*yes*) and without (*no*) comments. We report the F1-score, Kappa (K), and accuracy (Acc) for the positive and negative class.

6. DISCUSSION AND OUTLOOK

We presented exploratory work on predicting the comment volume of news articles prior to publication. We have developed a set of surface, cumulative, textual, semantic, and real-world features and report on their individual and combined performance on two classification tasks: Classify articles according to whether they will (i) generate comments, and (ii) receive few or many comments. Textual and semantic features prove to be strong performers, and the combination of all features leads to more robust classification.

To better understand our results, we look at misclassified instances. We identified five main types of error: (i) The event discussed in the news article is prone to comments, but this particular event is happening too far away (geographically). (ii) The event may be a comment "magnet," but is too local in this case. (iii) The news article itself is not attracting comments, but one posted comment sparks discussion. (iv) Shocking, touching, or in other ways surprising articles often generate more comments than can be expected from the article's content. (v) From the content of the article, a "controversial" topic might be expected, but the actual event is rather uncontroversial. Our failure analysis indicates that the features used in this paper are not the only factors involved in the prediction process. Future work should therefore focus on extracting more feature sets (e.g., context and entity-relations), use different encodings for current features, optimize the number of textual and semantic features per source, and explore optimized feature sets.

Acknowledgments. This research was supported by the DuOMAn project (STE-09-12) carried out within the STEVIN programme and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

7. REFERENCES

[1] D. L. Altheide. *Qualitative Media Analysis (Qualitative Research Methods)*. Sage Pubn Inc, 1996.
[2] D. S. Chung. Interactive features of online newspapers: Identifying patterns and predicting use of engaged readers. *J. Computer-Mediated Communication*, 13(3):658–679, 2008.
[3] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Can blog communication dynamics be correlated with stock market activity? In *HT '08*, pages 55–60. ACM, 2008.
[4] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05*. ACM, 2005.
[5] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *LA-WEB '07*, pages 57–66. IEEE Computer Society, 2007.
[6] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López. Homogeneous temporal activity patterns in a large online communication space. *CoRR*, abs/0708.1579, 2007.
[7] K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Coling 2008*, pages 473–480, 2008.
[8] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *AAAI-CAAW '06*, pages 145–152, 2006.
[9] G. Mishne and M. de Rijke. A study of blog search. In *ECIR '06*, pages 289–301, 2006.
[10] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *WWE '06*, 2006.
[11] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *WIDM '07*, pages 97–104. ACM, 2007.
[12] G. Szabó and B. A. Huberman. Predicting the popularity of online content. *CoRR*, abs/0811.0405, 2008.
[13] E. Tsagkias, M. Larson, and M. de Rijke. Exploiting surface features for the prediction of podcast preference. In *ECIR '09*, 2009.